Original Paper

Evaluating a Customized Version of ChatGPT for Systematic Review Data Extraction in Health Research: Development and Usability Study

Jayden Sercombe*, BPsych; Zachary Bryant*, BCom, MPH; Jack Wilson, PhD

The Matilda Centre for Research in Mental Health and Substance Use, University of Sydney, Sydney, Australia
*these authors contributed equally

Corresponding Author:

Zachary Bryant, BCom, MPH
The Matilda Centre for Research in Mental Health and Substance Use
University of Sydney
Jane Foss Russell Building (G02), Level 6
Sydney 2006
Australia
Phone: 612 8627 9380
Email: zachary.bryant@sydney.edu.au

Abstract

Background: Systematic reviews are essential for synthesizing research in health sciences; however, they are resource-intensive and prone to human error. The data extraction phase, in which key details of studies are identified and recorded in a systematic manner, may benefit from the application of automation processes. Recent advancements in artificial intelligence, specifically in large language models (LLMs) such as ChatGPT, may streamline this process.

Objective: This study aimed to develop and evaluate a custom Generative Pre-Training Transformer (GPT), named *Systematic Review Extractor Pro*, for automating the data extraction phase of systematic reviews in health research.

Methods: OpenAI's GPT Builder was used to create a GPT tailored to extract information from academic manuscripts. The Role, Instruction, Steps, End goal, and Narrowing (RISEN) framework was used to inform prompt engineering for the GPT. A sample of 20 studies from two distinct systematic reviews was used to evaluate the GPT's performance in extraction. Agreement rates between the GPT outputs and human reviewers were calculated for each study subsection.

Results: The mean time for human data extraction was 36 minutes per study, compared to 26.6 seconds for GPT generation, followed by 13 minutes of human review. The GPT demonstrated high overall agreement rates with human reviewers, achieving 91.45% for review 1 and 89.31% for review 2. It was particularly accurate in extracting study characteristics (review 1: 95.25%; review 2: 90.83%) and participant characteristics (review 1: 95.03%; review 2: 90.00%), with lower performance observed in more complex areas such as methodological characteristics (87.07%) and statistical results (77.50%). The GPT correctly extracted data in 14 instances (3.25% in review 1) and four instances (1.16% in review 2) when the human reviewer was incorrect.

Conclusions: The custom GPT significantly reduced extraction time and shows evidence that it can extract data with high accuracy, particularly for participant and study characteristics. This tool may offer a viable option for researchers seeking to reduce resource demands during the extraction phase, although more research is needed to evaluate test-retest reliability, performance across broader review types, and accuracy in extracting statistical data. The tool developed in the current study has been made open access.

JMIR Form Res 2025;9:e68666; doi: 10.2196/68666

Keywords: artificial intelligence; systematic reviews; data extraction; LLM; ChatGPT; AI; large language models

Introduction

The application of artificial intelligence (AI) in health research has the potential to innovate and optimize various research tasks [1]. Systematic reviews are now increasingly conducted as a gold-standard method for evaluating and synthesizing the evidence base [2]. However, this research methodology often takes up significant time, and therefore can be costly, and is subject to human error [3-5]. As large language models (LLMs) and their capabilities rapidly improve, they are now being used to assist with systematic reviews, streamlining a significant portion of health research methodology.

There has been a dramatic rise in the production of systematic reviews, as demonstrated by a 20-fold increase over the past 20-years, equivalent to 80 publications per day [6,7]. Although they have become a valuable resource for informing policy and clinical practice [8], systematic reviews require significant labor and financial costs. A recent analysis concluded that the mean estimated time to complete and publish a systematic review was 67.3 weeks, with an average of five authors per published text [3]. In addition to software and dissemination fees, organizations can expect to pay around USD $140,000 per review [4]. There is further concern for the potential of human error during the screening and extractions stages [5].

In response to this significant burden, there have been strong calls for the use of AI in systematic reviews, making the process more efficient [9]. The oldest applications have focused on the automation of title and abstract screening. Machine learning classifiers, such as support vector machines or complement naive Bayes have been trained to replicate human inclusion/exclusion decisions based on label training data [10,11]. Active learning methods, such as those implemented in softwares such as Covidence [12] or ASReview [13], prioritize the most relevant citations for initial screening. However, these traditional methods still require human involvement to make final decisions on study inclusion, and they do not automate other stages of the review (eg, extraction). Rule-based and natural language processing (NLP)–driven named entity recognition approaches have been used for structured data extraction from full-text articles [14]. Recent developments in deep learning, particularly transformer-based models such as BERT and SciBERT, have improved the performance of such tasks by capturing more nuanced linguistic features [15-17]. The recent advancement in NLP have further contributed toward the development of much-needed extraction tools [18-20]. While these pilot programs demonstrated strong performance, they were largely designed for the extraction of clinical trial data, limiting their application across study designs. Recent focus has shifted toward the popular platform, OpenAI's ChatGPT [21]. This accessible LLM software is proficient at understanding and processing human language with high speed and accuracy. As the tool can effectively interpret information from complex texts, it may have strong potential for application in the extraction of systematic reviews. In addition, OpenAI recently released the functionality to build customizable versions of ChatGPT. Users can create a Generative Pre-training Transformer (GPT) and tailor it for certain tasks by providing it instructions and assumed knowledge [22]. These applications can be saved and accessed by other users. With the recent GPT functionality, health researchers can now create and share tools specifically designed to carry out systematic review data extraction.

Despite recent advances in AI technology, only three studies to date have reported using LLMs for data extraction. One study briefly reported on the feasibility of ChatGPT for data extraction [23], while others evaluated the performance of ChatGPT-4 [24] and ChatGPT-3.5 [25] ; these studies found moderate performance for extracting complex information but high accuracy for simpler extraction fields [24,25]. While these studies are the first to evaluate the performance of ChatGPT as an extraction tool, further efforts are needed to improve the availability of these programs.

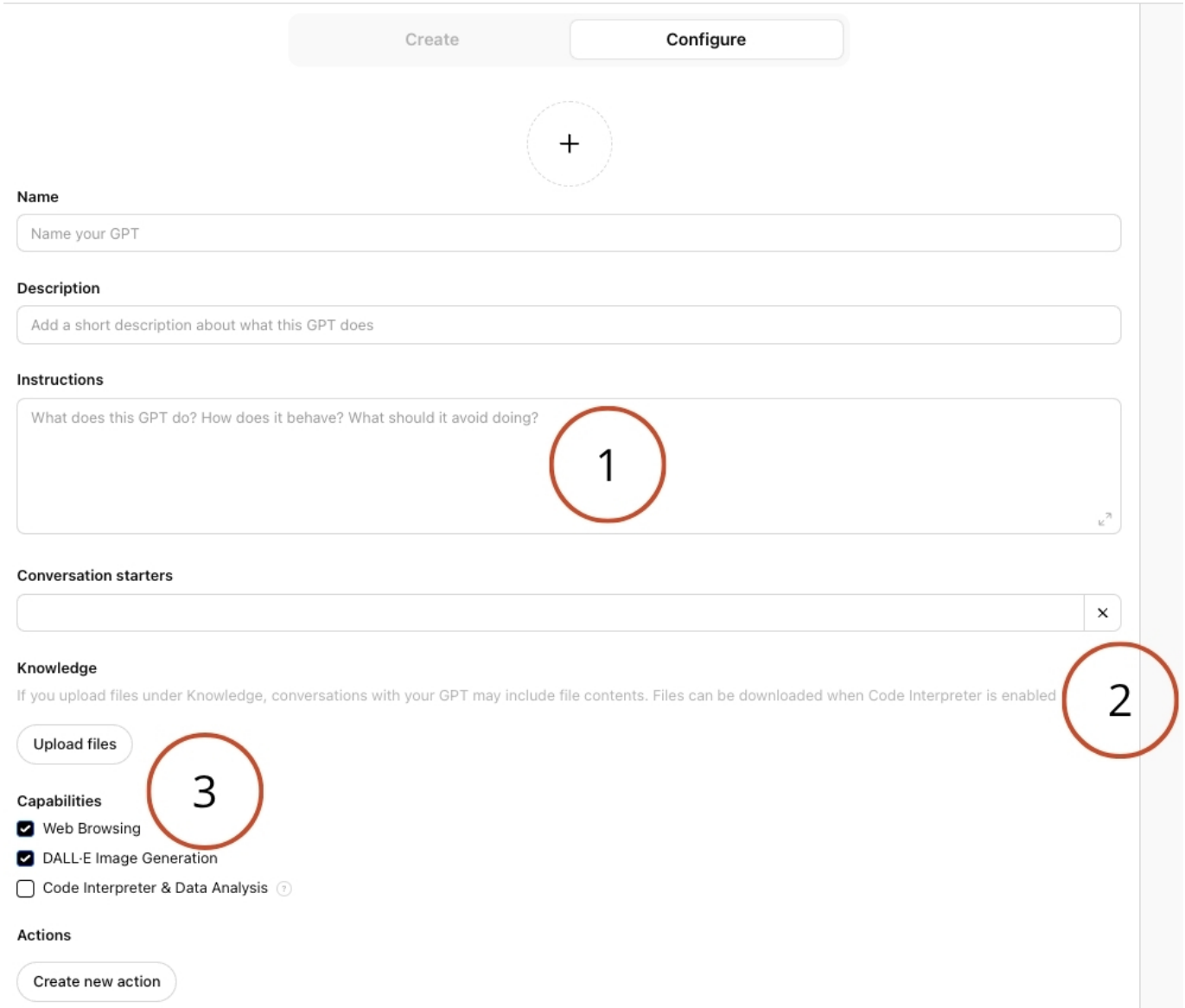
We present a custom GPT program tailored for systematic review extraction, made available as open access. In the current study, we aim to provide a pilot evaluation of the extraction tool, using data from two systematic reviews. Performance of the custom GPT will be compared to that of a human extractor, examining rates of agreement and time taken to extract data.

Methods

GPT Building

OpenAI's GPT (version 3.5) Builder allows users to create custom versions of ChatGPT that perform specific tasks by combining instructions, knowledge, and capabilities. The GPT builder interface was used to develop a specialized GPT to extract information from academic manuscripts to assist with the extraction phase of two systematic reviews: a methodological review and a systematic review of interventions. The developed GPT is named *Systematic Review Extractor Pro* (Figure 1).

Figure 1. The OpenAI custom GPT Builder interface used to develop and test automated data extraction (1–Instructions, 2–Knowledge, and 3–Capabilities).



Prompt Engineering

Detailed prompt engineering guidelines were followed under the Role, Instruction, Steps, End goal, and Narrowing (RISEN) framework [22] to manually configure the GPT's base context (purpose) and stepped actions to improve its performance in completing the required tasks. Each term refers to specific prompt components that improve the output quality. *Role* provides an identity to the GPT dictating the manner in which it acts, that is,"As an expert in the conduct of systematic reviews you are to...." *Instructions* inform the GPT what task it is to perform, while *Steps* provide a hierarchy of instructions that the GPT is to follow to perform the desired task. *End goal* informs the GPT on the format of desired output and content. Finally, *Narrowing* provides constraints to the GPT including key requirements for performing the given task (eg, setting a word limit on the GPT's response output).

Iterative Development of the Tool

The RISEN prompt used to instruct the GPT was iteratively fine-tuned to improve its performance by working through five iterations of the detailed instructions. Differences between iterations were generally focused on providing greater specificity to the prompt and detailed steps to carry out the extraction. Any vagueness in the *Steps* would lead to diverse response ranges for extraction variables, and therefore required the greatest number of changes and troubleshooting. Any changes made to the prompt were systematically tested on dummy health research manuscripts to optimize performance. Due to higher inaccuracy in early stages of extracting statistical results in the template, OpenAI's user guidelines were followed, to iteratively improve the tool [22]. These guidelines suggest 'splitting' or 'chunking' complex tasks into smaller ones.

Reviews for Extraction

A random sample of 10 studies from each of the two systematic reviews (total of 20 studies) were selected to evaluate the performance of the GPT. Review 1 is a systematic review of ecological momentary assessment (EMA) research methodologies for measuring substance use and associated behaviors. Full details are provided in the review protocol which was preregistered on the International Prospective Register of Systematic Reviews (PROSPERO: CRD42023400418). Of specific importance to this review was the extraction of methodological characteristics relating to ecological momentary assessment techniques. Review 2 is a systematic review of the effectiveness of wellbeing interventions for helping professionals. For comprehensive details, refer to the review protocol preregistered on PROSPERO (CRD42023422224). As a review assessing intervention effectiveness, this study included the extraction of statistical results.

Agreement

Total proportion of agreement was calculated by summing the number of instances where the GPT and human reviewer agreed on extracted data, divided by the total number of applicable fields. Agreement was computed for each study according to subsection, which included study characteristics, participant characteristics, methodological characteristics (only relevant for review 1), and statistical results (only relevant for review 2). See Multimedia Appendices 1 and 2 for the full list of variables in each subsection. A brief qualitative synthesis of the nature of errors was also conducted.

Timing

The time taken to extract data for each study was recorded for both the GPT and human reviewer. To enhance practicality, we also recorded the time taken for a human to review the GPT output.

Ethical Considerations

This study is a methodological paper which conducts secondary analysis on deidentified and aggregated data and as such, is exempt from requiring ethical approval as per the Australian National Statement on Ethical Conduct in Human Research 2023 [26].

Results

Time Taken

The mean extraction time for the human reviewers was 36 minutes. For GPT, the mean time for extraction was 26.6 seconds, and an additional 13 minutes for human review.

Agreement

For the completed extraction templates for the GPT and human reviewers, see Multimedia Appendix 1 (review 1) and Multimedia Appendix 2 (review 2). As seen in Table 1, overall agreement between the GPT and human reviewer across the 10 studies in review 1 was 91.45%; agreement was the highest for study (95.25%) and participant characteristics (95.03%) and the lowest for methodological characteristics (87.07%). For the 10 studies in review 2, overall agreement was 89.31%, similar to review 1. The highest agreement was observed in study (90.83%) and participant (90.00%) characteristics. Lower agreement was recorded in extraction fields relating to statistical results (77.50%).

Table 1. Agreement between human reviewers and GPT for data extraction by manuscript subsections.

Manuscript subsection	Review 1 (agreement %)	Review 2 (agreement %)
Study characteristics	95.25	90.83
Participant characteristics	95.03	90.00
Methodological characteristics	87.07	—[a]
Results	—[a]	77.50
Overall agreement	91.45	89.31

[a]Data not applicable for review.

Qualitative Summary of Errors

Across both Review 1 and 2, the most common error made by the GPT was reporting an extraction field as "not specified," despite the relevant information being present in the study manuscript (see Multimedia Appendices 1 and 2). These types of errors accounted for a high proportion of the inaccuracies, particularly where only p-values were sometimes reported rather than the required measures of effect.

While there were instances of errors resembling hallucination (eg, describing a randomized controlled trial as a pre-post study), these were few in number. When the GPT erroneously inserted information, it typically mistakenly drew data from other parts of the paper, rather than fabricating the information entirely, For example, several errors occurred in sample size extraction, such as reporting the baseline sample as the follow-up sample.

Instances Where the GPT Was More Accurate That the Human Reviewer

In review 1, there were 14 instances where the GPT correctly extracted fields, and the human reviewer was incorrect (equating to 3.25% of the overall number of fields). In Review 2, the GPT was correct in four fields where the human reviewer was incorrect with a rate of 1.16%.

Discussion

Principal Findings

This study evaluated the performance of a custom-configured LLM, specifically a GPT, in automating the data extraction phase of two systematic reviews. This application of AI for extraction provides initial insights into addressing challenges in systematic review methodology, such as time consumption, labor intensity, and potential for human error.

Even when accounting for human review, the custom GPT took less than 14 minutes to complete extraction of each study, 22 minutes faster than human extraction. The 26.6 seconds it took for extraction alone was comparable to a study by Gue et al [25], where ChatGPT-3.5 took approximately 17 seconds per study compared to 77 minutes by a reviewer. As such, the findings of prior research alongside the current study emphasize the processing speed of AI and its potential to significantly reduce the time demands and resultant costs associated with systematic reviews. This efficiency could facilitate more timely summaries of evidence-based practices and support researchers in managing large volumes of literature.

The high overall agreement rates observed in this study between the GPT and human extraction for review 1 (91.45% overall) and 2 (89.31%), suggest that LLMs may have potential to support automation of the extraction of data from academic manuscripts. Consistent across review 1 and 2, the GPT was particularly effective at extracting study (95.25% and 90.83%) and participant characteristics (95.05% and 90%, respectively). This is likely due to the homogeneity in participant and study characteristics that are reported in the extracted manuscripts, reflecting reporting standards upheld by academic journals.

However, agreement rates were lower when extracting methodological characteristics (77.88%) and results (77.50%) sections. Early iterations were improved by 'splitting' the task by extracting results separately to the rest of the extraction template. Results are challenging for GPT to extract, as studies used varied methods of analysis and some statistical results were only reported in tables. As LLMs favor extracting data from text information, additional prompting was required to ensure that tables were thoroughly and consistently incorporated into the information extraction sequence.

In this study, a small percentage of instances where the GPT outperformed the human extractor were observed. This demonstrates the potential for AI to identify patterns and insights that may be overlooked in human extraction. Human extractors can be prone to fatigue which may have led to these small inconsistencies. The preliminary findings of this study suggest that LLMs may have utility in identifying errors made by human extractors.

Overall, in accordance with previous research, LLMs excel in quickly extracting simple data (eg, participant demographics [25]. The purpose-built GPT evaluated in the current study was less accurate while extracting more complex fields compared to simple fields (agreement ranged from 77.50% to 95.25%) but appears to be relatively accurate in comparison to extraction using the generic ChatGPT model where agreement rates were as low as 41.2% [25]. A study of 87 human reviewers found error rates of 28.3% to 31.2%, corresponding to agreement rates of 68.8% and 71.7% with a gold-standard reviewer [27]. These rates remained relatively consistent across reviewer experience level and, similar to the findings of the current study, were more accurate for the extraction of participants and study characteristics compared to statistical results. Overall, the GPT evaluated in the current study demonstrated comparable error rates to those reported by human reviewers.

In its current format, researchers may consider the use of this tool as a second extractor, when they lack funding, resourcing, or time for a team member to perform the task (Textbox 1). The overall agreement rate of the custom GPT was approximately 90%, and the GPT had utility in identifying some human errors. Caution should be exercised while using the tool to extract information in areas where it showed lower accuracy, in particular results and methodological characteristics.

Textbox 1. A guide to using the Systematic Review Extractor Pro for data extraction.

How to use the tool

1. Log into a ChatGPT account and access the 'Systematic Review Extractor Pro' GPT [28].
2. Copy and paste your extraction template variable headings into the GPT message bar (any text format).
3. Upload a PDF file of the study from which you want to extract data.
4. Copy the output to an Excel file. If needed, use the 'Convert Text to Table' functionality, separating by 'l'.

Optimizing use of this tool

- Be explicit with the outcomes in your extraction template. For example, "Mental health outcomes (specify scales and subscales used in bullet point format)" or "modality (in-person; digital; telephone)." For an example, see the authors templates in Multimedia Appendices 1 and 2.
- With more complex aspects of the template such as statistical results, you may have better accuracy by 'chunking' the task. This involves splitting complex tasks into simpler ones. To do this, provide the tool with the results section of the extraction template separate to the rest of the template.

Citing this tool and other considerations

- To reference the use of this tool in publications, please cite this paper.
- Consider the copyright limitations of articles when uploading them to the tool.

Strengths and Limitations

Building on previous work exploring the capacity of ChatGPT 3, 3.5 and 4 to perform data extraction [23-25], this study tested the performance of a custom-built GPT across two distinct systematic reviews. To the best of our knowledge, this study highlights the first empirical investigation of a custom OpenAI GPT to perform data extraction in a systematic review.

Our study has several limitations. When determining agreement between extractors, there is always a degree of subjectivity between coding any qualitative responses, which can either overestimate or underestimate agreement rates. The current study observed agreement rates across 20 studies. Testing a larger sample size would improve confidence in the tool. In terms of limitations of the tool itself, the tool's construction on the GPT Builder may pose replicability challenges of the output due to the constantly evolving nature of the LLM. The benefit of this aspect is that the tool will automatically update as ChatGPT releases improved versions of their model [29]. However, future research should evaluate the test-retest validity of this custom GPT.

Conclusions

This study serves as promising early evidence for the application of a custom-built GPT in conducting systematic review data extraction. The findings of this study are consistent with the growing sentiment suggesting AI can enhance the efficiency of systematic reviews and reduce cost and time while maintaining accuracy [4]. By encouraging open discourse regarding the role of AI in research, researchers can contribute to the development of robust, transparent, and reproducible scientific practice. This is a rapidly advancing field of technology and OpenAI frequently releases new versions of ChatGPT that show improved performance and efficiency [29]. It is likely that with further advancement of AI, LLMs may be relied on as a sole reviewer for data extraction.

Acknowledgments

JS and ZB are supported by Postgraduate Scholarship through the University of Sydney. JW is supported by a NHMRC fellowship funding. While this manuscript describes the use of generative AI as a tool for systematic review data extraction, the authors attest that no generative AI was used in the writing of this manuscript.

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Authors' Contributions

Conceptualization: JB, JS
Formal analysis: JB, JS
Investigation: JS
Writing–original draft: JB, JS, JW
Writing–review & editing: JS, JW

Conflicts of Interest

None declared.

Multimedia Appendix 1

Systematic review 1 - agreement file.
[XLSX File (Microsoft Excel File), 27 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Systematic review 2 - agreement file.
[XLSX File (Microsoft Excel File), 46 KB-Multimedia Appendix 2]

References

1. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med. Jan 2022;28(1):31-38. [doi: 10.1038/s41591-021-01614-0] [Medline: 35058619]
2. Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ. Welch VA, editor. Cochrane Handbook for Systematic Reviews of Interventions Version 65. Cochrane; 2024. URL: https://www.training.cochrane.org/handbook [Accessed 2024-09-15]
3. Borah R, Brown AW, Capers PL, Kaiser KA. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. BMJ Open. Feb 27, 2017;7(2):e012545. [doi: 10.1136/bmjopen-2016-012545] [Medline: 28242767]
4. Michelson M, Reuter K. The significant cost of systematic reviews and meta-analyses: a call for greater involvement of machine learning to assess the promise of clinical trials. Contemp Clin Trials Commun. Dec 2019;16:100443. [doi: 10.1016/j.conctc.2019.100443] [Medline: 31497675]
5. Mathes T, Klaßen P, Pieper D. Frequency of data extraction errors and methods to increase data extraction quality: a methodological review. BMC Med Res Methodol. Nov 28, 2017;17(1):152. [doi: 10.1186/s12874-017-0431-4] [Medline: 29179685]
6. Hoffmann F, Allers K, Rombey T, et al. Nearly 80 systematic reviews were published each day: observational study on trends in epidemiology and reporting over the years 2000-2019. J Clin Epidemiol. Oct 2021;138:1-11. [doi: 10.1016/j.jclinepi.2021.05.022] [Medline: 34091022]
7. Johnson BT, Hennessy EA. Systematic reviews and meta-analyses in the health sciences: best practice methods for research syntheses. Soc Sci Med. Jul 2019;233:237-251. [doi: 10.1016/j.socscimed.2019.05.035]
8. Chalmers I, Fox DM. Increasing the incidence and influence of systematic reviews on health policy and practice. Am J Public Health. Jan 2016;106(1):11-13. [doi: 10.2105/AJPH.2015.302915] [Medline: 26562111]
9. van Dijk SHB, Brusse-Keizer MGJ, Bucsán CC, van der Palen J, Doggen CJM, Lenferink A. Artificial intelligence in systematic reviews: promising when appropriately used. BMJ Open. Jul 7, 2023;13(7):e072254. [doi: 10.1136/bmjopen-2023-072254] [Medline: 37419641]
10. O'Mara-Eves A, Thomas J, McNaught J, Miwa M, Ananiadou S. Using text mining for study identification in systematic reviews: a systematic review of current approaches. Syst Rev. Jan 14, 2015;4(1):5. [doi: 10.1186/2046-4053-4-5] [Medline: 25588314]
11. Wallace BC, Small K, Brodley CE, Trikalinos TA. Active learning for biomedical citation screening. Presented at: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining; Jul 25, 2010:173-182; Washington DC USA. [doi: 10.1145/1835804.1835829]
12. Covidence systematic review software. 2025. URL: https://www.covidence.org/ [Accessed 2024-09-09]
13. van de Schoot R, de Bruin J, Schram R, et al. ASReview: Open source software for efficient and transparent active learning for systematic reviews. Int J Digit Curation. 2021;16(1):123-133. [doi: 10.2218/ijdc.v16i1.726]
14. Jonnalagadda SR, Goyal P, Huffman MD. Automating data extraction in systematic reviews: a systematic review. Syst Rev. Dec 2015;4(1):78. [doi: 10.1186/s13643-015-0066-7]
15. Beltagy I, Lo K, Cohan A. SciBERT: a pretrained language model for scientific text. Presented at: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); Nov 7-10, 2019:3615-3620; Hong Kong, China. [doi: 10.18653/v1/D19-1371]
16. Khalil H, Ameen D, Zarnegar A. Tools to support the automation of systematic reviews: a scoping review. J Clin Epidemiol. Apr 2022;144:22-42. [doi: 10.1016/j.jclinepi.2021.12.005] [Medline: 34896236]
17. van de Schoot R, de Bruin J, Schram R, et al. An open source machine learning framework for efficient and transparent systematic reviews. Nat Mach Intell. Feb 2021;3(2):125-133. [doi: 10.1038/s42256-020-00287-7]
18. Golinelli D, Nuzzolese AG, Sanmarchi F, et al. Semi-Automatic systematic literature reviews and information extraction of COVID-19 scientific evidence: description and preliminary results of the COKE project. Information. 2022;13(3):117. [doi: 10.3390/info13030117]
19. Panayi A, Ward K, Benhadji-Schaff A, Ibanez-Lopez AS, Xia A, Barzilay R. Evaluation of a prototype machine learning tool to semi-automate data extraction for systematic literature reviews. Syst Rev. Oct 6, 2023;12(1):187. [doi: 10.1186/s13643-023-02351-w] [Medline: 37803451]

20. Zhang T, Yu Y, Mei J, Tang Z, Zhang X, Li S. Unlocking the power of deep PICO extraction: step-wise medical NER identification. arXiv. 2020. URL: http://arxiv.org/abs/2005.06601 [Accessed 2024-09-10]
21. ChatGPT. Open AI. 2023. URL: https://openai.com/chatgpt/overview/ [Accessed 2025-08-07]
22. Prompt engineering. Open AI. 2024. URL: https://platform.openai.com/docs/guides/prompt-engineering [Accessed 2024-07-03]
23. Mahuli SA, Rai A, Mahuli AV, Kumar A. Application ChatGPT in conducting systematic reviews and meta-analyses. Br Dent J. Jul 2023;235(2):90-92. [doi: 10.1038/s41415-023-6132-y] [Medline: 37500847]
24. Khraisha Q, Put S, Kappenberg J, Warraitch A, Hadfield K. Can large language models replace humans in systematic reviews? Evaluating GPT-4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages. Res Synth Methods. Jul 2024;15(4):616-626. [doi: 10.1002/jrsm.1715] [Medline: 38484744]
25. Gue CCY, Rahim NDA, Rojas-Carabali W, et al. Evaluating the OpenAI's GPT-3.5 Turbo's performance in extracting information from scientific articles on diabetic retinopathy. Syst Rev. May 16, 2024;13(1):135. [doi: 10.1186/s13643-024-02523-2] [Medline: 38755704]
26. Council and Universities Australia. National Statement on Ethical Conduct in Human Research. National Health and Medical Research Council; 2023. URL: https://www.nhmrc.gov.au/about-us/publications/national-statement-ethical-conduct-human-research-2023 [Accessed 2024-09-10]
27. Horton J, Vandermeer B, Hartling L, Tjosvold L, Klassen TP, Buscemi N. Systematic review data extraction: cross-sectional study showed that experience did not increase accuracy. J Clin Epidemiol. Mar 2010;63(3):289-298. [doi: 10.1016/j.jclinepi.2009.04.007] [Medline: 19683413]
28. Systematic review extractor pro. ChatGPT. URL: https://chatgpt.com/g/g-5IRhGHEcc-systematic-review-extractor-pro [Accessed 2025-08-07]
29. OpenAI, Achiam J, Adler S, Agarwal S, et al. GPT-4 technical report. arXiv. Mar 4, 2024. URL: http://arxiv.org/abs/2303.08774 [Accessed 2024-09-10]

Abbreviations

AI: artificial intelligence
GPT: Generative Pre-Training Transformer
LLM: large language model
RISEN: Role, Instruction, Steps, End goal, and Narrowing

Edited by Amaryllis Mavragani; peer-reviewed by Ashutosh Kumar Dubey, Jiaping Zheng; submitted 11.11.2024; final revised version received 01.07.2025; accepted 03.07.2025; published 11.08.2025

Please cite as:
Sercombe J, Bryant Z, Wilson J
Evaluating a Customized Version of ChatGPT for Systematic Review Data Extraction in Health Research: Development and Usability Study
JMIR Form Res 2025;9:e68666
URL: https://formative.jmir.org/2025/1/e68666
doi: 10.2196/68666